Chapter 15
Interactive Statistical Modeling with XGms

Jean-Bernard Martens

Abstract This chapter intends to clarify how statistical models can be used in the interactive analysis of experimental data. It is discussed how maximum likelihood estimation can map experimental data into model parameters, and how confidence intervals on these parameters play a key role in drawing inferences. One frequent inference is whether or not similar data collected in two different experimental conditions require a significant change in model parameters, hence indicating that the conditions are not equivalent. Another frequent inference is whether a simple or a more complex model needs to be adopted for describing observed data. It is argued that the user, who has insight into the problem underlying the data, should be allowed to play an active part in selecting models and specifying the inferences of interest. Hence, the need for interactive visualization is identified and the program XGms that has been developed for this purpose is discussed. The fact that the statistical models implemented within XGms are geometrical in nature simplifies their visualization and interpretation.

Keywords: Statistical modeling, Interactive modeling, Model visualization, Data analysis

15.1 Introduction

There is a widespread interest in establishing and understanding relationships between variables, where a variable is defined as any measurable aspect that can take on different values. Sometimes this interest is motivated by sheer curiosity, in which case it can suffice to establish whether or not there is a relationship between the variables of interest. In industrial or scientific settings, however, the interest is often triggered by practical or theoretical necessity, and more quantitative descriptions of relationships are usually required.

An established way of testing and describing relationships between variables is by means of statistical analysis. Statistics can aim at drawing inferences about whether or not variables are related, such as in the case of a T-test or chi-squared

E. Zudilova-Seinstra et al. (eds.) *Trends in Interactive Visualization*,
doi:10.1007/978-1-84800-269-2_15

test [13], or can aim at estimating quantitative relationships between observed data, such as in the case of regression. In order to argue the need for more intuitive interactions in statistical data analyses, we highlight a number of characteristics of current practice.

First, a statistic is a single number derived from the data, and detailed characteristics of the data, such as individual data points that do not satisfy a global relationship, can be obscured by such statistics. Statistics should therefore be used primarily to verify effects that can also be visualized through other means, such as with scatter or parallel plots [20] or confidence intervals [12, 19]. Therefore, most popular statistical packages, such as SPSS, Statgraphics, DataDesk, etc., offer interactive visualizations for both exploratory data analysis (i.e., the stage in which hypotheses are construed and statistical methods are selected) and for illustrating the results of statistical analyses. The required analyses themselves however most often need to be specified using command-line entry or menu selection, which requires a substantial familiarity with the available methods, their properties and their limitations.

Second, in order to compress a large set of data into a single number (or a few numbers), assumptions need to be made about the data being interpreted. For example, the widely used analysis of variance (ANOVA) assumes that the observed data is normally distributed across conditions, with varying average, but constant variance. Being able to rapidly switch between alternative assumptions, such as changing to non-constant variance, and to adequately visualize the impact of such changes, is essential for gaining insight into the validity of such assumptions. Currently, such an exploration of alternatives cannot be performed in an intuitive way. Selecting alternative statistical methods and finding the relevant information in the wealth of textual and graphical output provided by these methods is what is now most often required. This is a slow and non-intuitive way of working.

Third, a plethora of statistical methods have been developed in order to cope with alternative data characteristics (such as whether the data is discrete or continuous, interval or ordinal) and with different hypotheses (such as, testing for equal averages or variances). Because of this complexity, a statistical expert may be needed to help select the most adequate method. Since this is seldom feasible, most users tend to restrict themselves to the most simple and best-known methods, even in cases where such methods are clearly not the best choice (or even simply inappropriate). Although the underlying mathematical models for many of the existing methods are more closely related than most people seem to be aware of, end users are not able to perceive these communalities. A better visualization of the models being available for approximating the data can potentially lead to choosing more appropriate methods, which in turn is likely to result in better abstractions from the data.

In this chapter, we propose that interactive visualization can help to make statistical data modeling more intuitive and appropriate. More specifically, we propose to enable the user to interactively control different aspects of a general statistical model in order to explore a range of alternative model descriptions, rather then having to select from a list of available statistical methods. Moreover, we propose an output visualization that is closely related to this general model so that the consequences of alternative model choices can be easily appreciated. We also propose that the user,

who has insight in the problem underlying the observed data, should be allowed to play an active role in the model estimation process by providing reasonable guesses, hence avoiding local optimizations. Such interactive feedback and feed-forward is deemed crucial when developing an understanding of observed data.

In Sect. 15.2 of this chapter, we introduce some fundamentals of statistical modeling, including maximum likelihood estimation and (asymptotic) confidence intervals. In Sect. 15.3, we discuss a class of geometrical models that are both intuitive and flexible, i.e., able to model data with very different characteristics. The majority of this section is dedicated to illustrating how the model can be used to address a diversity of statistical problems, such as ordinal and linear regression, factor analysis and multidimensional scaling. In Sect. 15.4, we describe the interface of XGms, the interactive visualization program that implements these geometrical models. We end up by identifying important contributions and aspects for future development.

15.2 Statistical Modeling

15.2.1 Single Data, Single Model

A statistical model describes the *probability* of any possible outcome of an experiment. More precisely, $\boldsymbol{x}$ let be the vector of observations and let $P(\boldsymbol{x}; \theta)$ be the probability of this observation vector in case the vector of model parameters is equal to θ. Once the actual observation $\boldsymbol{x}$ is known, $P(\boldsymbol{x}; \theta)$ can be viewed as a function of the parameter vector θ. This function is called the *likelihood function* (LF) for the model, given the observation vector $\boldsymbol{x}$. Let us illustrate this concept by means of a simple example. Suppose that the experiment consists of N independent observations of a binary output, and that the number of positive (1) and negative (0) outcomes is n and $N - n$, respectively. The outcome of such an experiment could be modeled by a binomial distribution with probability p for a positive outcome, i.e.,

$$P(n, N-n; p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \tag{15.1}$$

This corresponds to assuming that all observations are performed independently and that p is constant across observations. Since n and N are known from the observation, the LF only depends on the model parameter p.

From all possible models, we want to select the one that best describes our actual data vector $\boldsymbol{x}$. According to the *maximum likelihood* (ML) principle, the optimum choice θ for the parameter vector θ is such that the likelihood function $P(\boldsymbol{x}; \theta)$ is maximized. Since probabilities are often exponential functions, it is customary to maximize the (natural) logarithm of the LF, i.e., $L(\theta) = \log P(\boldsymbol{x}; \theta)$, as a function of θ. Once the ML parameter estimate $\hat{\theta}$ is known, then we can construct the log likelihood profile (LLP) $\Lambda(\theta) = 2[L(\theta) - L(\hat{\theta})]$. In case of our simple binomial model, the log likelihood function (LLF) is equal to

$$L(p) = \log \frac{N!}{n!(N-n)} + n \log p + (N-n)\log(1-p) \tag{15.2}$$

Maximizing this function with respect to the parameter p, i.e.,

$$\frac{dL(p)}{dp} = \frac{n}{p} - \frac{N-n}{1-p} = 0, \tag{15.3}$$

results in the optimal ML estimate $\hat{p} = n / N$, and a LLP equal to

$$\Lambda(p) = 2N(\hat{p} \log \frac{\hat{p}}{p} + (1-\hat{p}) \log \frac{1-\hat{p}}{1-p}) \tag{15.4}$$

This LLP is plotted in Fig. 15.1a for two possible experimental observations with the same optimal ML estimate $\hat{p} = n / N$, but different total number N of observations. Note that the minimum becomes more pronounced as the number of observations N increases, and that the parabolic approximation to the LLP (to be defined below) also improves.

Intuitively, we expect that more extensive data sets, with a higher number of observations, lead to more accurate predictions for the model parameters than smaller data sets. Within statistical modeling, this is formalized through the concept of

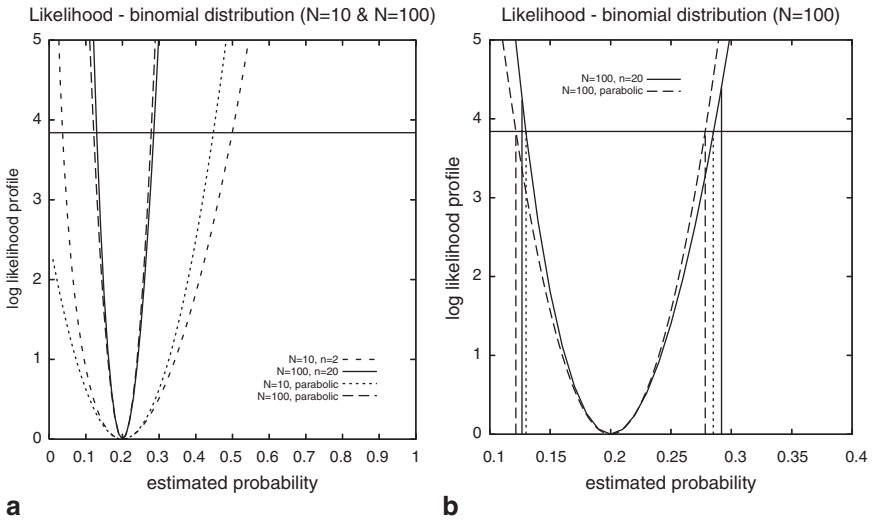

Fig. 15.1 Log likelihood profiles and their parabolic approximations for a binomial model with one (probability) parameter p. Figure (**a**) shows the model for two different observation vectors with identical ML estimate, $\hat{p} = 0.2$ i.e., (2, 8) (N = 10) and (20, 80) (N = 100). An enlarged view for observation (20, 80) is shown in Fig. (**b**). The vertically drawn lines in this figure indicate the exact 95% confidence interval for the model parameter p. The dashed and dotted vertical lines show two different, asymptotically correct, approximations. The horizontal line is the $\chi^2_{0.05}(1) = 3.84$ boundary

confidence intervals (CIs). More precisely, we expect the ML parameter estimate $\hat{\theta}$ to vary slightly in case of repeated experiments. This is captured by the statistical nature of the model, since a model with constant parameter vector θ will give rise to different observations $\boldsymbol{x}$ on successive (simulated) trials. Without going into detail (see [14, 17] for more in-depth treatments), a $100(1-\alpha)\%$ CI is a region in parameter space, surrounding the ML estimate $\hat{\theta}$, such that the "true" model parameter vector θ is expected to fall within this region with an estimated probability of $100(1-\alpha)\%$ (the most frequently adopted choice is 95%, or $\alpha = 0.05$). The hypothesis that a model with assumed parameter vector θ_o describes the data can then be rejected with $100(1-\alpha)\%$ confidence if this vector falls outside the $100(1-\alpha)\%$ CI. Exact determination of these CIs is only possible in relatively simple special cases, such as the binomial model discussed above [6]. The exact 95% CI for the parameter p in case of an (20, 80) observation is indicated by the drawn vertical lines in Fig. 15.1b. Since $p_o = 0.3$ falls outside of the 95% CI, we can reject the hypothesis that the probability p underlying our observation is $p = 0.3$ (we can for instance not reject the hypothesis that $p = 0.25$).

The popularity of ML estimates can be largely attributed to the fact that there exist approximations to the CIs that can be derived in very general cases and that are asymptotically correct, i.e., in case the number of observations is (infinitely) large. Provided the number of parameters in the model is k (this is the number of elements in the parameter vector θ), then the "likelihood ratio" approximation [14] to the CI is obtained by

$$\Lambda(\theta) < \chi_a^2(k) \tag{15.5}$$

where $\chi_\alpha^2(k)$ is such that the area from $\chi_\alpha^2(k)$ to infinity under the chi-squared distribution curve with k degrees of freedom is equal to α (which means that the area under the curve between 0 and $\chi_\alpha^2(k)$ is equal to $1-\alpha$). These chi-squared values are tabulated in almost all textbooks on statistics, e.g., [4, 13]. The "likelihood ratio" approximation to the 95% CI for the parameter p in the binomial model of Fig. 15.1b is indicated by the dotted vertical lines. The boundary values for the interval are obtained by intersecting the LLP $\Lambda(p)$ with the horizontal line at height $\chi_{0.05}^2(1) = 3.84$. Note that the asymptotic CI (ACI) is (slightly) smaller than the true CI.

In case the LLP around the optimum ML estimate $\hat{\theta}$ can be approximated by a parabolic function, which is guaranteed to be the case when the number of observations is large (due to the central limit theorem), then the CI can be shown to be approximately equal to an ellipsoid with expression

$$(\theta - \hat{\theta})^t \cdot I(\hat{\theta}) \cdot (\theta - \hat{\theta}) = \chi_\alpha^2(k), \tag{15.6}$$

where the k x k matrix of the negated second derivatives to the log likelihood function, i.e.,

$$I(\hat{\theta}) = \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} L(\hat{\theta}); i, j = 1,...k \right] \tag{15.7}$$

is called the observed *Fisher information matrix* [14]. For the binomial model with one parameter p, this matrix reduces to a single number

$$I(\hat{p}) = -\left.\frac{d^2 L(p)}{dp^2}\right|_{p=\hat{p}} = \frac{N}{\hat{p}\,(1-\hat{p})} \tag{15.8}$$

which corresponds to a parabolic or "Wald" approximation to the LLP, i.e.,

$$\hat{\Lambda}(p) = \frac{N}{\hat{p}(1-\hat{p})}(p-\hat{p})^2 \tag{15.9}$$

The "Wald" approximation to the 95% CI for the parameter p in the binomial model is indicated by the dashed vertical lines in Fig. 15.1b. The boundary values for the interval are obtained by intersecting the parabolic approximation $\hat{\Lambda}(p)$ to the LLP with the horizontal line at height $\chi^2_{0.05}(1) = 3.84$.

15.2.2 Multiple Data and/or Models

In the above discussion we have assumed that there is only one observation vector $\boldsymbol{x}$ and one model of interest, with model parameter vector θ. In practice, the situation is usually more complex in at least two respects.

First, we often have different observation vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_K$ that correspond to different experimental conditions, and the question of interest is whether or not the experimental conditions have a significant effect on the observations. In order to compare such observations, we can model them using an identical model with varying parameters. More specifically, if the observation vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_K$ are modeled by different ML estimated parameter vectors $\hat{\theta}_1,\ldots,\hat{\theta}_K$ within a common model, then the statistical question (or hypothesis) of interest is whether or not the model parameters $\hat{\theta}_i$ and $\hat{\theta}_j$, or a selected subset of them, are significantly different. Since many different comparisons are possible, interactive visualization can be used to select only the interesting ones, and to adequately visualize the comparison results.

Second, we might be interested in testing different models, of increasing complexity, on the same observation vector $\boldsymbol{x}$. If the first model has k_θ parameters, and the second model is an extension of the first one with $k_\varphi > k_\theta$ parameters, then the latter model is expected to have a higher log likelihood (LL). The statistical question of interest is knowing whether or not this increase in LL is sufficiently substantial to warrant the increase in number of parameters, in which case the more complex model should be preferred. If this increase in LL is not substantial enough, then the simpler model should be preferred. Exploring such alternative models in an efficient way will be another objective of using interactive visualization.

In order to address the first aspect, we will need a slightly more general formulation of ML estimation than the one provided before. More specifically, we will need to know how to determine CIs for subsets of parameters. Let us divide the parameter vector $\theta = (\theta_1, \theta_2)$ into two subsets of parameters (with k_1 and k_2 components, respectively, so that $k_1 + k_2 = k$) and rewrite the LLP as $\Lambda(\theta) = \Lambda(\theta_1, \theta_2)$. It has been demonstrated [14] that the LLP for the "wanted" parameter vector θ_1 equals

$$\Lambda(\theta_1) = \Lambda(\theta_1, \hat{\hat{\theta}}_2(\theta_1)), \tag{15.10}$$

where $\hat{\hat{\theta}}_2(\theta_1)$ is the "unwanted" parameter vector that maximizes $\Lambda(\theta_1, \theta_2)$ for a fixed value of θ_1. This log likelihood profile tends to a χ^2-distribution with k_1 degrees of freedom for large data sets, so that an $100(1 - \alpha)\%$ ACI for the subset θ_1 of "wanted" parameters can be obtained by setting

$$\Lambda(\theta_1) < \chi^2_\alpha(k_1) \tag{15.11}$$

Alternatively, a parabolic approximation [14]

$$\tilde{\Lambda}(\theta_1) = (\theta_1 - \hat{\theta}_1)^t \cdot (I_{11} - I_{12} I_{22}^{-1} I_{21}) \cdot (\theta_1 - \hat{\theta}_1) \tag{15.12}$$

to the log likelihood profile can be obtained, where the matrices I_{ij}, of size $k_i \times k_j$, for $i,j = 1,2$, are derived by subdividing the original observed Fisher matrix of size $k \times k$, see equation (15.7), in sub-matrices, i.e.,

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \tag{15.13}$$

We now demonstrate how this result can be used to derive an ACI for the difference between model parameters. Assume that the same model, with different model parameters θ_i and θ_j, is used to model the observation vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively. If both experiments are performed independently, then the probability for the combined observation vector $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the product $P(\boldsymbol{x}_i, \theta_i)$. $P(\boldsymbol{x}_j, \theta_j)$, so that the LLF is simply

$$L(\theta_i, \theta_j) = \log P(\boldsymbol{x}_i; \theta_i) + \log P(\boldsymbol{x}_j; \theta_j); \tag{15.14}$$

where the probability function $P(\boldsymbol{x}; \theta)$ is identical for both observation vectors. The following coordinate transformation

$$\delta_{ij} = \frac{\theta_i - \theta_j}{2}, \theta_i = \rho_{ij} + \delta_{ij} \tag{15.15}$$

$$\rho_{ij} = \frac{\theta_i + \theta_j}{2}, \theta_j = \rho_{ij} - \delta_{ij}$$

can be used to express this LLF $L(\theta_i, \theta_j) = L(\delta_{ij}, \rho_{ij})$ in terms of two new parameter vectors (δ_{ij}, ρ_{ij}). The procedure in the previous paragraph tells us how to derive an ACI for the difference vector δ_{ij}. We can reject the hypothesis that both model vectors θ_i and θ_j are equal if the zero vector lies outside of this ACI. The procedure is easily adapted in case we only want to compare a subset of the model parameters. More precisely, by further subdividing the difference vector $\delta_{ij} = (\delta_{ij}(1), \delta_{ij}(2))$ into two subsets, we can apply the above procedure to create an ACI for the difference vector $\delta_{ij}(1)$ with the parameters of interest (removing both $\delta_{ij}(2)$ and θ_{ij} as unwanted parameters).

We now address the second aspect mentioned above, i.e, that of selecting between increasingly complex models for the same observation vector $\boldsymbol{x}$. More specifically, assume that the LLFs are $L_\theta(\mu) = \log P_\theta(\boldsymbol{x};\ \theta)$ for the simpler model with k_θ parameters, and $L_\phi(\phi) = \log P_\phi\ (\boldsymbol{x};\ \phi)$ for the more complex model with k_ϕ parameters. The probability functions are now different, since the two models are different. It has been demonstrated [17] that the log likelihood difference (LLD)

$$\Delta\Lambda = 2 \cdot [L_\phi(\hat{\phi}) - L_\theta(\hat{\theta})] \tag{15.16}$$

where $\hat{\theta}$ and $\hat{\phi}$ are the optimal ML estimates for the model parameters, is asymptotically χ^2-distributed with $k_\phi - k_\theta$ degrees of freedom. Hence, only in case this LLD exceeds the threshold $\chi^2_\alpha(k_\phi - k_\theta)$ do we consider adopting the more complex model description over the simpler one. Note that, in practice, there frequently arise situations where we prefer to stick to a simpler model that is easier to interpret, despite the statistical argument to the contrary.

15.3 Geometrical Models for Statistical Data Analysis

Now that we have established how parameters of statistical models can be estimated from observations and know how ACIs can be used to draw inferences between alternative models, or model parameters, we can direct our attention to a specific class of statistical models.

According to the principle of *homogeneity of perception* [7], different variables associated with a set of objects are often related, and the relationships may be visualized by linking them to geometrical properties of a stimulus configuration. More specifically, distances between points representing objects can be related to observed (dis)similarities between these objects, while coordinates of object points along different axes can be associated with measurements on different attributes. Such models provide a geometrical interpretation for popular statistical methods such as multiple regression (MR) and multidimensional scaling (MDS) [1, 3, 7, 11]. In case of MR, part of the model (i.e., the object positions) is a priori provided, based on some theoretical or physical understanding of the objects of interest. In case of MDS, these object positions are estimated from the experimental data.

In order to clarify the above concepts, we illustrate part of the geometrical model that is implemented within our program XGms in Fig. 15.2. We assume that N different objects, represented by object positions $\boldsymbol{x}_i$, for $i = 1,\ldots,N$, have been assessed on K_a different attributes, and that measurements have been repeated R_a times, so that A_{kij} is the j-th measurement, for $j = 1,\ldots,R_a$, of object i on attribute k, for $k = 1,\ldots,K_a$. According to the model, the object configuration $\{\boldsymbol{x}_i;\ i = 1,\ldots,N\}$ is shared by all attribute measurements. The mapping of an object position $\boldsymbol{x}_i$ to a response A_{kij} for attribute k depends on the prediction vector $\boldsymbol{a}_k$, the proportionality factor f_k, the offset factor c_k, the noise standard deviation σ_{ak} and the possibly nonlinear but monotonically increasing response function $R_{ak}(\cdot)$, for $k = 1,\ldots,K_a$. How

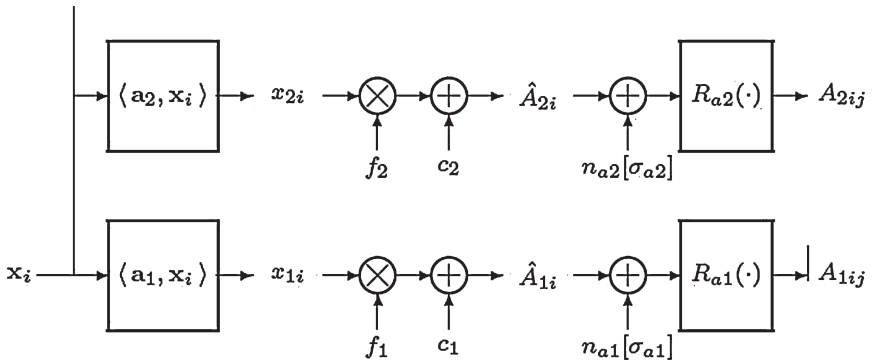

Fig. 15.2 Geometrical model relating object positions $\boldsymbol{x}_i$ to measurements A_{kij} for variable k on object i, during repetition j. The objects within an experiment are described by positions $\boldsymbol{x}_i$, while the attribute vector $\boldsymbol{a}_k$, regression coefficients c_k and f_k, and response mechanism $R_{\alpha k}$ describe the mapping to the variable measurements on these objects. The response variability in successive trials j around the predicted mean value $\hat{A}_{ki}$ for variable k on object i is modeled by a noise source n_{ak} with (constant) standard deviation $\sigma_{\alpha k}$

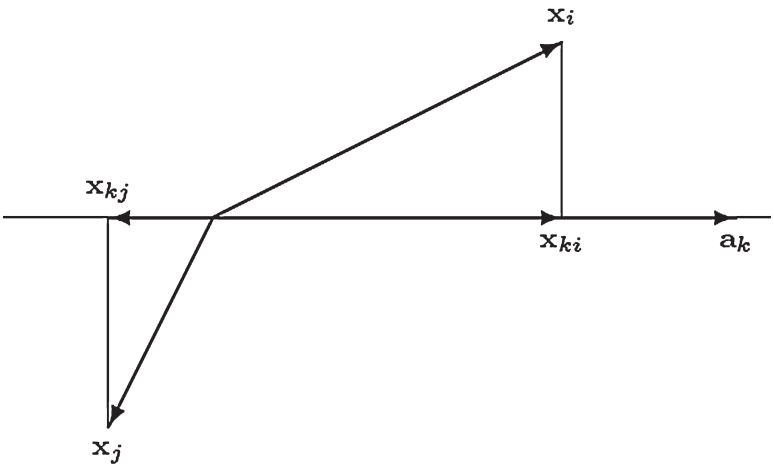

Fig. 15.3 Geometrical interpretation of attribute predictions. If $\boldsymbol{x}_{ki}$ is the orthogonal projection of the vector $\boldsymbol{x}_i$ onto the axis with direction $\boldsymbol{a}_k$, then the inner product definition implies that $< \boldsymbol{a}_k, \boldsymbol{x}_i > = < \boldsymbol{a}_k, \boldsymbol{x}_{ki} >$. The projected vector can be expressed as $\boldsymbol{x}_{ki} = \boldsymbol{x}_{ki} \cdot \boldsymbol{a}_k / ||\mathrm{a}_k||_2$ since it is aligned with the vector a_k. The numbers $\boldsymbol{x}_{ki}$ and $\boldsymbol{x}_{kj}$ can be interpreted as coordinates on the axis with direction $\boldsymbol{a}_k$ for two different objects i and j

object position $\boldsymbol{x}_i$ is mapped into an (average) score x_{ki} for object i on variable k is illustrated in Fig. 15.3. When the same objects are scored very differently on different attributes, then this is reflected within the model as attribute vectors $\boldsymbol{a}_k$ with different orientations. Such distinct orientations are obviously only possible in

case the dimension n of the space exceeds one. Note that the model in Fig. 15.2 is statistical in nature because of the inclusion of the (Gaussian) noise sources n_{ak}. For a detailed mathematical description of how the probabilities of observed data values A_{kij} are related to the model parameters, we refer to [10, 11].

The proposed geometrical model is especially flexible in how object positions are mapped into attribute measurements. In this way, observed data with very different characteristics can be handled within a common model, and different alternative assumptions about the data characteristics can be easily compared. More specifically, the response function $R_{ak}(\cdot)$ can be linear in case of *interval* (or metric) data, but is also allowed to be a non-linear, monotonically increasing function, in case of *ordinal* (or non-metric) data. Both power-like functions [15] and monotonically increasing spline functions [16] are supported. The response function can optionally be quantized, i.e., limited to a discrete set of output values, in case of *discrete* data, such as in case of a measurement variable with integer instead of real values [10]. Variables with different ranges and accuracies can be handled by adapting the linear regression coefficients c_k and f_k and the noise standard deviations σ_{ak}.

15.3.1 Example 1. Binary Data

We start with a simple illustrative example where the stimulus configuration is one-dimensional, i.e., each condition under test is represented by a single number x_i, for $i = 1,\ldots,N$. We assume that a constant number of $R_a = 100$ binary (positive/negative) measurements are performed in each of $N = 19$ conditions, and that the fraction of negative outcomes in condition i is $p_i = i*0.05$, for $i = 1,\ldots,19$. Since the same binary measurement is performed in each condition, the number of measured attributes is $K_a = 1$. Such data can be analyzed using the model of Fig. 15.2 by adopting a nonlinear mapping $R_{a1}(\cdot)$ that quantizes the output to two levels, representing the positive and negative outcomes. A small fraction of negative outcomes in case i should map to an average prediction $\hat{A}_{1i}$ well above the quantization threshold of the nonlinearity (corresponding to a small probability of generating a value below the threshold), while an equal fraction of negative and positive outcomes should map to a position close to the quantization threshold.

The estimated positions for the different conditions are plotted in Fig. 15.4. For reasons explained in [10], the stimulus configuration is always centered around the origin and its variance is normalized. The estimation program determines that the offset c_1 in the model is equal to the threshold value of the quantizing nonlinearity $R_{a1}(\cdot)$, which defaults to 0, and that the scale factor f_1 is equal to 0.864 times the noise standard deviation $\sigma_{\alpha 1}$, which defaults to 1. This implies that the standard deviation of the noise on the stimulus configuration in Fig. 15.4 equals $\sigma_{a1}/f_1 = 1.157$.

During the ML estimation, only the stimulus configuration is visualized, because calculating ACIs at each iteration step would significantly slow down the estimation process. The ACIs are moreover only required when inferences need to be drawn from the data. Therefore, Fig. 15.4 is only generated in response to an

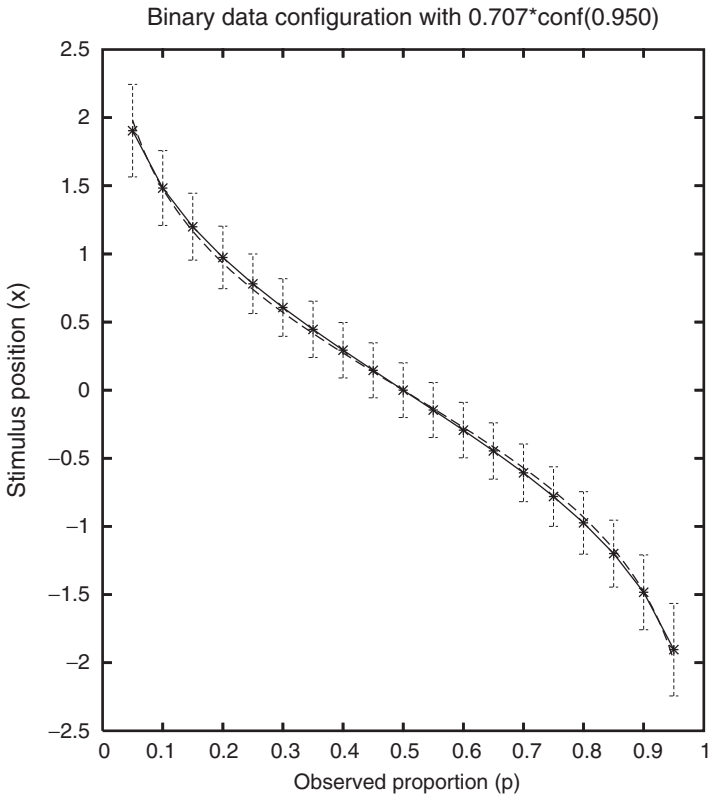

Fig. 15.4 Estimated stimulus positions for binary data with a constant decrease in observed fraction p of positive responses. The intervals equal $1/\sqrt{2}$ times the 95% CIs on the estimated positions. The dashed curve is proportional to the logarithm of the odds ratio $\log((1-p)/p)$

explicit user request. Although extensive statistical tests (following the ML principles discussed in the previous section) can be reported textually, the figure being presented is a deliberate approximation that is much easier to interpret. Indeed, if all $N(N-1)/2 = 171$ stimulus distances $x_i - x_j$ would need to be reported and checked for statistical significance, a long list that is difficult to inspect would be the result. Of course, we could prune this list by only reporting the results for neighboring stimulus positions. Instead, we have adopted an approximate graphical rendering that has recently been promoted by several authors [12, 19]. It has been shown that, provided the standard errors on the estimated positions are approximately constant (an assumption that is reasonably satisfied in our case), we can perform a simple visual inspection in order to identify statistically significant differences. More precisely, if we render $1/\sqrt{2}$ times the estimated 95% CI around each position, then non-overlapping intervals are equivalent to statistically significant differences. Note for instance that observed proportions 0.05 and 0.15 are significantly different according to this criterion (which is in agreement with both the exact Fisher test [5], with $p = 0.0317$, or the approximate χ^2-test [13], with $p = 0.018$), while observed

proportions 0.15 and 0.25 are not (which is in agreement with both the exact Fisher test, with $p = 0.111$, or the approximate χ^2-test, with $p = 0.077$). Note that the intervals between stimulus positions x_i can be compared on a linear scale (because of the assumption of constant noise variance), which is not allowed for the observed fractions p_i themselves. This is one of the purposes of building geometrical models. If the noise would be logistically rather than normally distributed (which is only a minor difference), then theoretical considerations lead to the conclusion that the stimulus configuration should be proportional to the logarithm of the odds ratio, i.e., $x_i \alpha \log((1 - p_i)/p_i)$ [6]. Figure 15.4 shows that this is also a good approximation in case of a Gaussian noise model.

15.3.2 Example 2. One-Dimensional Regression

Linear regression [4] is undoubtedly the most popular method for studying associations between independent and dependent variables when the latter variable is continuous and metric. More specifically, linear regression tests whether or not there is evidence for a linear relationship between both variables. The model in Fig. 15.2 can simultaneously test the linear regression between one independent variable, assigned to the positions x_i, for $i = 1,..,N$, and multiple (K_a) dependent variables. In order to keep the discussion simple we will restrict ourselves to the simple case of only one dependent variable. More specifically, we will consider example data from Table 24.2 in reference [4], which is reproduced in Table 15.1.

Table 15.1 Measured concentration of chlorine in a product as a function of the number of weeks after production (Reproduced from [4], Table 24.2)

Weeks	Chlorine concentration
8	0.49, 0.49
10	0.48, 0.47, 0.48, 0.47
12	0.46, 0.46, 0.45, 0.43
14	0.45, 0.43, 0.43
16	0.44, 0.43, 0.43
18	0.46, 0.45
20	0.42, 0.42, 0.43
22	0.41, 0.41, 0.40
24	0.42, 0.40, 0.40
26	0.41, 0.40, 0.41
28	0.41, 0.40
30	0.40, 0.40, 0.38
32	0.41, 0.40
34	0.40
36	0.41, 0.38
38	0.40, 0.40
40	0.39
42	0.39

The model of Fig. 15.2 describes linear regression between the dependent variable A_{1ij} (observed chlorine concentration within a medicine) and the independent variable x_i (number of weeks since production) in case the response function is linear, i.e., $R_{a1}(A) = A$ (with inverse $T_{a1}(A) = R_{a1}^{-1}(A) = A$). Within the program XGms, the result of the regression is presented in a scatter plot, where the regressed values $\hat{A}_{1i} = c_1 + f_1 \cdot x_i$ are used as coordinates along the horizontal axis, and the transformed measured values $T_{a1}(A_{1ij})$ are used as coordinates along the vertical axes (see Fig. 15.8). In case of linear regression, the data points should align along the diagonal line that is included as a visual aid. An alternative would be to plot residuals $T_{a1}(A_{1ij}) - \hat{A}_{1i}$ [4] instead of observed measurements $T_{a1}(A_{1ij})$ along the vertical axis. In order to assess the goodness of fit, two lines that run parallel to the diagonal at a distance of two times the estimated noise standard deviation σ_{a1} are also included. In case of an adequate model fit, approximately 95% percent of the data points should fall in between these oblique lines. In case of our data set, a reasonable model fit was observed, but with some systematic deviations. In order to complement the visual impression, the user can request more detailed information about the regression. The result is presented in both graphical and textual form. In the graphical output, the 95% ACI for the linear regression coefficient f_1 is plotted. Since zero was outside this confidence interval, it could be concluded that there was evidence for a linear relationship. The complementary textual output contains a complete Analysis of Variance (ANOVA). Since there were repeated measures, ANOVA could also perform an F-test for the lack-of-fit [4]. This test, which is very often omitted, establishes whether or not there are variations within the dependent data that are not explained by the linear regression. The result $F(16,26) = 5.201$ ($p = 0.00011$) > 2.052($p = 0.05$) indeed indicates that there is a lack of fit. We discuss two options for how to (interactively) obtain more insight into this deviation from a linear regression.

The first option is to perform non-linear instead of linear regression. Within the model of Fig. 15.2, this corresponds to including a nonlinear response function R_{a1}. Increasingly complex response functions can be created, up to the point where increasing the number of parameters no longer contributes significantly to improving the regression. The LLD $\Delta\Lambda$ of equation (15.16) can be used to judge significance, since the XGms program can be requested to compare the current model with a previously stored model at any time during the interactive session. Two kinds of nonlinear transformations are supported by XGms, power-like transformations with up to 3 parameters, and monotonic spline transformations with up to 5 internal knot points, which corresponds to 6 parameters [10]. In both cases, only the first two parameters add significantly to improving the (nonlinear) regression. The optimal power-law and spline transformations are both presented in Fig. 15.5a. The optimal nonlinear transformation is also displayed during the interactive visualization (see Fig. 15.8), so that the observer gets a visual impression of it (for instance useful in judging the deviation from linearity).

The second option for analyzing the non-linearity is to compare the independent data with the optimal linear predictor. Within the model of Fig. 15.2, this corresponds to MDS optimization of the configuration $\boldsymbol{x}_i$ based on the observed

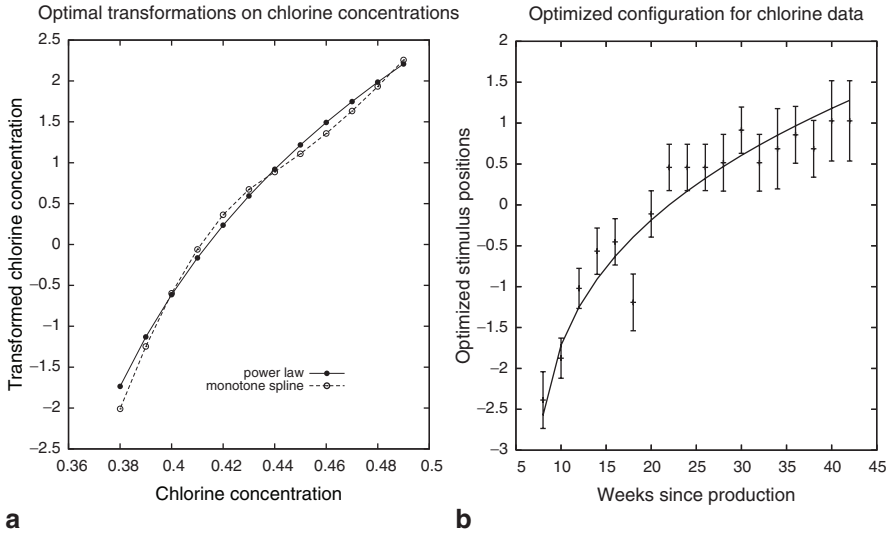

Fig. 15.5 Figure (**a**) shows optimized power-law and monotonically increasing spline transformations on chlorine concentrations. Figure (**b**) is an optimized (MDS) stimulus configuration with approximate linear relationship to the dependent variable (chlorine concentration), plotted as a function of the independent variable (weeks since production). The drawn line is an approximate power-law transformation with two parameters on the independent variable

data, and to plotting the optimized configuration as a function of the independent variable. The result of this MDS analysis is shown in Fig. 15.5b. As already observed in the previous section, the MDS stimulus configuration is geometrically (i.e., linearly) related to the dependent variable in case there is no output non-linearity. The observed relationship in Fig. 15.5b hence indicates which transformation needs to be applied to the independent variable (the weeks since production) to have an approximate linear effect on the dependent variable (the chlorine concentration). The difference with the first method, where the transformation was applied to the dependent variable, is that the second method determines the transformation that is required on the independent variable. The model in Fig. 15.2 does not allow for a transformation at this stage (i.e., on the scores x_{ki}, before they enter the linear regression). Nevertheless, the model can be used to find a suitable approximation. Indeed, by keeping the optimized MDS configuration fixed while changing the dependent data to what was originally the independent data (i.e., the weeks since production), one can apply the regression method of the previous section to find an optimum non-linear transformation on the independent variable. The optimum two-parameter power-law transformation derived in this way is also shown in Fig. 15.5b.

15.3.3 Example 3. Multiple Regression and Factor Analysis

The simple case treated in the previous example, where only one independent variable characterizes the objects under study, has its obvious limitations. In practice, the objects of interest will often vary in multiple aspects and the purpose of an experiment may be to establish a suitable compromise between conflicting attributes. This requires establishing how one or more dependent variables are (linearly or non-linearly) related to a number of independent parameters. With reference to the model in Fig. 15.2, this means that each object i under study, for $i = 1,\ldots,N$, is characterized by a vector $\boldsymbol{x}_i$, the elements of which are the independent variables. For these objects, one or more dependent variables may be measured. We will illustrate this case by an example where the perceived quality of images degraded by noise and blur is modeled [8, 10]. All combinations of four levels of blur with a binomial kernel and four levels of Gaussian additive white noise were applied to create 16 different processed images from an original image. Seven subjects rated the perceived quality of the images. All images were presented $R_a = 4$ times, in random order, to each of the subjects, resulting in $16 \times 4 = 64$ quality scores per subject. Integer scores between 0 and 10 were used by the subjects to express their judgments.

The multiple regression (MR) vectors ($f_k \cdot \boldsymbol{a}_k$ in the notation of Fig. 15.2) that describe the quality judgments of the individual subjects are plotted as dotted vectors in Fig. 15.6. The 4×4 stimulus configuration is square because the blur and the noise are assumed to be independent characteristics. For only one of the seven subjects was there evidence that nonlinear regression was significantly better than linear regression, so that linear regression was adopted for all judgments. Note that the 95%-confidence ellipses on the regression vectors (which are scaled by a factor of $1/\sqrt{2}$ in order to facilitate interpretation) indicate that some vectors have approximately the same orientation, while others have significantly different orientations. Establishing the number of independent factors underlying a set of attribute judgments is usually accomplished through mathematical procedures such as factor analysis. However, the interactive visualization within XGms also allows to gather evidence for which attributes share a common factor. More specifically, the number of parameters in the model of Fig. 15.2 can be reduced by letting some attribute vectors share the same orientation (i.e., by equating normalized attribute vectors, e.g., $\boldsymbol{a}_1 = \boldsymbol{a}_2$). Note that such vectors can still have unequal lengths (since f_1 need not be equal to f_2). If two attributes share the same orientation then the number of degrees of freedom in the model is reduced (by one in the case of a two-dimensional configuration), at the cost of a decrease in the log likelihood of the model.

The LLD of equation (15.16) can be used to decide whether or not this decrease is sufficiently large to prefer the more complex model, with unequally oriented attribute vectors, over the simpler model, with equally oriented attribute vectors. Within the interactive visualization program XGms, the user can select which attribute orientations to merge. For the specific data set considered in this example, an iterative procedure where the two most closely related orientations (i.e., with

smallest mutual angle) are merged in each successive step, leads to the conclusion that there is statistical evidence for only two groups of quality vectors (i.e., two independent factors). The resulting dashed vectors are also shown in Fig. 15.6. The vectors in a group share the same orientation, but have different lengths. The obvious interpretation for these two groups is that one group (of four subjects) attributes more relative weight to blur than to noise in comparison to the other group (of three subjects). It is of course possible that we are not interested in individual or group quality judgments, but only in the average quality direction across subjects. This is easily accomplished by letting all vectors share the same orientation ($\boldsymbol{a}_1 = \ldots = \boldsymbol{a}_7$). The resulting average orientation for quality corresponds to the drawn vector in Fig. 15.6.

15.3.4 Example 4. Multidimensional Scaling

The original motivation for developing the XGms program was to create an interactive visualization for multidimensional scaling (MDS). Within the model of Fig. 15.2, MDS refers to cases where the stimulus positions $\boldsymbol{x}_i$ are estimated, together with the other parameters, from the observed data A_{kij}. We can for instance use the data from the previous example to perform MDS. The resulting optimum 2D configuration

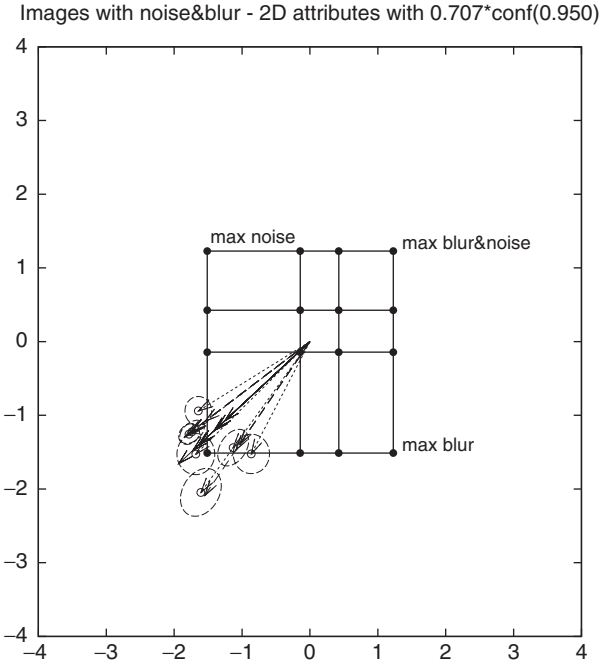

Fig. 15.6 MR model for quality judgments on images with blur and noise

is shown in Fig. 15.7, together with the regressed quality vectors (both the individual and group quality vectors, as well as the averaged quality orientation). This MDS solution has a significantly higher log likelihood than the multiple regression solution in the previous section ($\Delta\Lambda = 278.04 > \chi^2_{0.05}(28) = 41.35$), so that the MDS configuration of Fig. 15.7 is more accurate for predicting image quality than the MR configuration of Fig. 15.6. However, note the elongated confidence ellipses on the stimulus points, which indicate that the stimulus positions are ill determined in the direction orthogonal to the quality vectors. This is an instance where it would be interesting to compare the ellipsoidal ACIs with the "likelihood ratio" ACIs (something that is currently not implemented in XGms), since the elliptical approximation to the log likelihood may not be sufficiently accurate in this case. Comparison between a 2D and a 1D configuration using the log likelihood difference measure reveals that there is nevertheless statistical evidence ($\Delta\Lambda = 92.72 > \chi^2_{0.05}(16) = 26.31$) for a 2D configuration over a 1D configuration. Despite the fact that there is statistical evidence for a 2D solution, a 1D solution might be preferred because of the obvious difficulty of interpreting the configuration in Fig. 15.7. The problem is obviously due to the fact that the quality judgments only measure stimulus variations in one direction. It has been shown in [9, 10] that a much more stable 2D configuration arises when noise and blur judgments, which correspond to different directions in this space, are combined with the quality judgments.

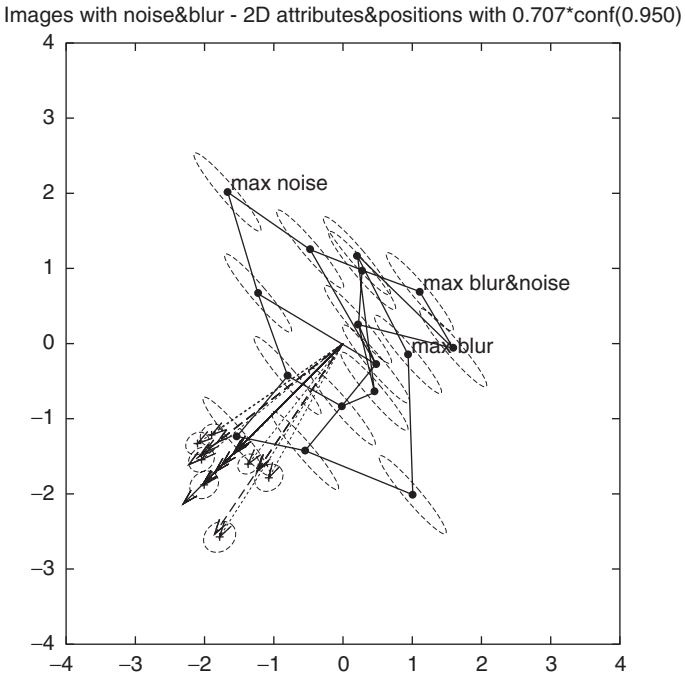

Fig. 15.7 MDS model for quality judgments on images with blur and noise

15.4 The XGms Interface to the Geometrical Model

In order to control model parameters and options, and in order to feedback the modeling results to the user, an appropriate interface to the geometrical model of Fig. 15.2 is required. Most existing statistical programs can be characterized as off-line programs (see for instance [1] for an overview of computer programs for MDS). This implies that the user must enter his/her choice of statistical method, and that the results are returned as text and graphs, to be examined after the analysis has finished. This makes it very difficult and cumbersome to appreciate the impact of alternative model choices, and hence to choose the model that most adequately describes the data. Also, since such programs are nonlinear optimization programs, the reported solution may correspond to a local optimum instead of a global one. The interactive program XGvis [2] is one of the few examples that has been developed to help remedy such problems. The program XGvis allows to interactively control the main parameters in an MDS model, and exchanges the calculated stimulus configuration with a second program, called XGobi [18].

This latter program is an interactive dynamic data visualization tool for the X Window environment. By combining the functionality of both programs, the user is not only able to dynamically alter the parameters of an MDS model within XGvis, but he can also view and manipulate the stimulus configuration in XGobi. This interactivity for instance allows the user to assist the optimization algorithm in avoiding sub-optimum solutions that correspond to local minima.

Like most MDS programs [1], the XGvis program only models continuous dissimilarity data of a single subject. In our view, this functionality is too limited for modeling more complex measurement situations, such as when relationships between multiple variables need to be established. In order to overcome these limitations, XGvis has been extended to include the joint analysis of data from single-stimulus scaling (such as in Fig. 15.2), double-stimulus difference scaling and dissimilarity scaling for multiple variables (the latter options are discussed in [10, 11]). Another extension of the resulting program XGms is that it not only supports continuous data but also discrete (metric and non-metric) data. The graphical user interface to XGms is depicted in Fig. 15.8 for the (Chlorine) data set in Example 15.2 of the previous section. This interface evolved from the graphical user interface of the XGvis program [2].

The panel at the top contains the major action buttons. The user can either specify an initial stimulus configuration at startup or can load a new stimulus configuration (using “Load CONF”) at any time during the XGms session. The current stimulus configuration can be used for MR analysis against the experimental data. Alternatively, the current stimulus configuration can be used as the initial configuration in an MDS analysis. Either the original data (Ai), power-transformed data (At) or spline-transformed data (As) can be used in the MR or MDS analysis. In the example, power-transformed attribute data are used. The stimulus configuration is continuously exchanged between the statistical modeling program XGms and the visualization program XGobi [18], so that this latter program can be used to manipulate and render the configuration. In order to allow

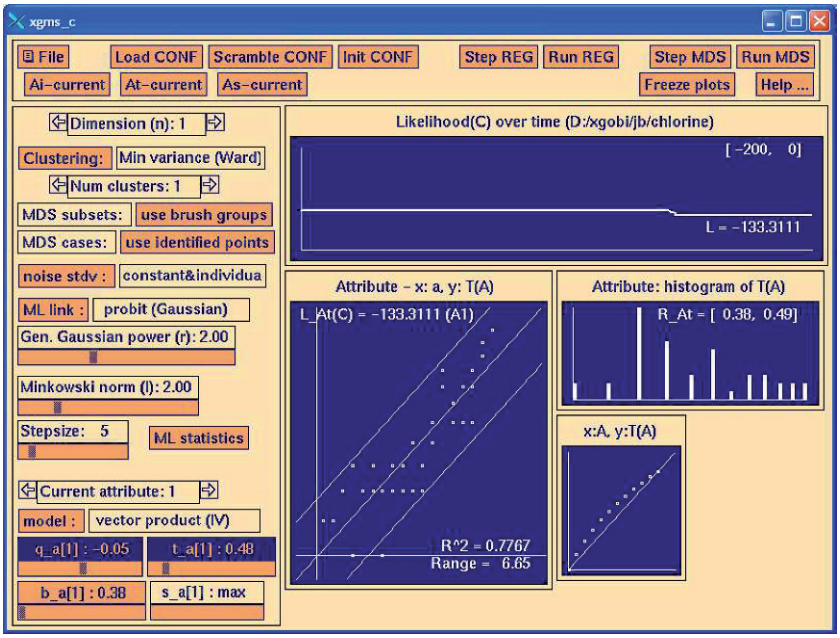

Fig. 15.8 XGms graphical user interface to the data set in Example 2 (*See Color Plates*)

processing by other data analysis or visualization programs, most intermediate data in the XGms program can also be output to ASCII files using the options in the "file" menu. File menu options can also be used to request more detailed graphical representations that include the ACIs needed for statistical inference.

The left panel in the interface allows the user to control the most important parameters of the geometrical model, such as

1. the dimension n of the stimulus space,
2. the noise model used,
3. the prediction model used (individual or shared attribute vector),
4. the parameters q, t and b that control the nonlinear power-like transformations on the variable data, and
5. the number s of internal knot points in a monotonically increasing spline transformation.

Xgms can also perform clustering on the stimulus configuration; and the left panel contains two action buttons that allow to interactively control this clustering. The user can obtain ML statistics, including the current number of DOF in the model and the likelihood value of the model, at any time in the XGms session by hitting the "ML statistics" button. XGms compares the outcomes on two successive calls to "ML statistics." In this way, χ^2-tests between alternative models can be made in order to infer which model is most appropriate. If data for several attributes are available, then the "current" selection allows to view the settings and model fits

for the current attribute. The user can for instance select which parameters of the power-like transformation are fixed to a user-specified value, and which can be optimized by XGms. Attributes can also be grouped so that they share a common prediction vector. It is also possible to exclude the subset of the data currently being viewed from the model optimization.

The graph in the upper-right corner displays how the ML optimization criterion (which is the negated log likelihood) has varied as a function of time during the XGms session. In addition, a scatter plot of transformed experimental data (y:T(A)) versus predicted model data (x:a), and a bar plot representing the histogram of the transformed experimental data (T(A)) are shown for the `current' attribute. The data in the scatter plot "Attribute - x:a, y:T(A)" can also be output to the visualization program XGobi in order to examine them more closely. The small panel "x:A, y:T(A)" depicts the monotonically increasing transformation that is performed in case the data are considered to be non-metric. It is the inverse $T_{ak}(a) = R_{ak}^{-1}(A)$ of the response function in Fig. 15.2.

15.5 Conclusions

We have argued how statistical modeling can profit from interactive visualization. Moreover, we have introduced a specific class of geometrical models that we consider to be intuitive, but nevertheless general enough to capture a large number of interesting data analysis situations. We have developed the interactive visualization program XGms for interaction with this class of models. Although the mathematical basis for this work was already presented in earlier publications [10, 11], experience with other users than the author has revealed that some aspects of the model are fairly easy to understand and apply, but that exploiting the full potential requires an introduction to the less straightforward aspects. This was an important motivation for the introduction to ML statistics and the discussion of example applications provided in this chapter.

We propose that an interactive visualization system for statistical modeling should contain three main components. First, it should incorporate a statistical modeling engine (or expert system) that can perform ML estimations of statistical model parameters, and that can construct the CIs required for statistical inference. Second, it should provide an interface through which users can intuitively formulate their statistical questions. Third, it should render the results of the analysis in a way that can be easily and quickly assessed by the users. Further developments in each of these components within XGms are foreseen. Concerning component one, we should note that the model shown in Fig. 15.2 covers only part of the modeling capabilities of the interactive visualization program XGms. More precisely, next to individual scaling data, the model can also handle pairwise comparisons, such as dissimilarity data or difference data [10]. Another important aspect is that currently only ellipsoidal ACIs (using the "Wald" approximation) are available, while there is an obvious interest to also implement ACIs based on the "likelihood ratio" approximation. Concerning component two, it is clear that extensive usage scenarios,

such as the ones described in the examples, are instrumental to establishing the kind of user actions that are potentially most useful. Additional support, such as for factor analysis of attribute vectors, could make the interaction at this stage less elaborate at times (currently, the reported factor analysis in Example 3 involved some text editing of input files and restarting of the program). A more intuitive interface than the current buttons and menus, i.e., one providing a more explicit relationship to the statistical model (such as in Fig. 15.2), would also be a great improvement, especially for first-time users. Inspiration might be drawn from related recent programs for Partial Least Squares modeling, i.e., Smart PLS (http://www.smartpls.de) and PLS-Graph (http://disc-nt.cba.uh.edu/plsgraph/), on how models can be graphically represented in an easy-to-understand way. The interface that is currently implemented within XGms (and XGobi) is heavily influenced by the options available within the low-level X Windows toolkit being used, and is clearly outdated. More advanced rendering techniques could obviously help to make the interface more appealing and revealing (for instance, better 3D rendering could lower the threshold for moving from 2D to 3D models).

15.6 Summary

This chapter explains how statistical models can be used in the interactive analysis of experimental data. It is discussed how maximum likelihood estimation can map experimental data into model parameters, and how confidence intervals on these parameters play a key role in drawing inferences. One frequent inference is whether or not similar data collected in two different experimental conditions require a significant change in model parameters, hence indicating that the conditions are not equivalent. Another frequent inference is whether a simple or a more complex model needs to be adopted for describing observed data. It is argued that the user, who has insight into the problem underlying the data, should be allowed to play an active part in selecting models and specifying the inferences of interest. Hence, the need for interactive visualization is identified and the program XGms that has been developed for this purpose is discussed. The fact that the statistical models implemented within XGms are geometrical in nature simplifies their visualization and interpretation. The flexibility of modeling diverse data in XGms is demonstrated by means of several examples.

It is proposed that an interactive visualization system for statistical modeling should contain three main components. First, it should incorporate a statistical modeling engine (or expert system) that can perform maximum likelihood estimations of statistical model parameters, and that can construct the confidence intervals required for statistical inference. Second, it should provide an interface through which users can intuitively formulate their statistical questions, such as which models to compare. Third, it should render the results of the analysis in a way that can be easily and quickly assessed by the users. The paper discusses how the program XGms meets with these demands and how it can possibly be extended in the future to overcome current limitations.

References

1. Borg I 0020, Groenen P (2005) Modern Multidimensional Scaling. Springer, New York
2. Buja A, SwayneD, LittmanM, Dean N, Hoffman H (2002) Visualization methodology for multidimensional scaling. Journal of Classification19, 7–44
3. Cox T, Cox M (1994) Multidimensional Scaling. Chapman & Hall, London
4. Draper N, Smith H (1998) Applied Regression Analysis. John Wiley and Sons Inc., New York
5. Fisher R (1992) On the interpretation of $\chi 2$ from contingency tables, and the calculation of p^2. Journal of the Royal Statistical Society 85, 87–94
6. Fleiss J, Levin B, Cho Paik M (2003) Statistical methods for rates and proportions. John Wiley and Sons Inc.,Hoboken, New Jersey
7. GreenP, Carmone Jr. F, Smith S(1998)Multidimensional Scaling, Concepts and Applications. Allyn & Bacon, Boston, Massachusetts
8. Kayargadde V, Martens JB (1996)Perceptual characterization of images degraded by blurand noise: Experiments. Journal of the Optical Society of America A13, 1166–1177
9. Kayargadde V, Martens JB (1996) Perceptual characterization of images degraded by blur and noise: Model. Journal of the Optical Society of America A13, 1178–1188
10. Martens J (2003) Image Technology Design – A Perceptual Approach.Kluwer Academic Publishers, Boston, Massachusetts
11. Martens JB (2002) Multidimensional modeling of image quality. Proceedings of the IEEE 90, 133–153
12. Masson M, Loftus G (2003) Using confidence intervals for graphically based data interpretation. Canadian Journal of Experimental Psychology 57, 203–220
13. Longman, an imprint of Addison Wesley Longman, Inc. New York – Reading, Massachusetts – Menlo Park, California – Harlow, England – Don Mills, Ontario – Sydney – Mexico City – Madrid – Amsterdam
14. Meeker W, Escobar L (1995)Teaching about approximate confidence regions based on maximum likelihood estimation.The American Statistician 49, 48–53
15. Ramsay J (1977) Maximum likelihood estimation in multidimensional scaling. Psychometrika 42,241–266
16. Ramsay J (1977) Monotonic weighted power transformations to additivity. Psychometrika 42,83–109
17. Stuart A, Ord J (1991) Kendall's Advanced Theory of Statistics: Volume 2 – Classical Inference and Relationship, 5th edn. Edward Arnold, London
18. Swayne D, Cook D, Buja A (1998) Xgobi: Interactive dynamic data visualization in the x window system. Journal of Computational and Graphical Statistics 7, 113–130
19. Tryon WW (2001) Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. Psychological Methods 6(4), 371–386
20. Unwin A, Volinsky C, Winkler S (2003) Parallel coordinates for exploratory modeling analysis. Computational Statistics & Data Analysis 43(4), 553–564